STCF 值:基于研究主题的学术文献影响力评价新指标*

李秀霞 宋凯
曲阜师范大学传媒学院 日照 276826

摘要:[目的/意义]学术文献影响力评价指标不断推陈出新,但尚缺乏在研究主题层面对文献影响力的评价,为发现不同研究主题内具有高影响力和引用价值的文献,本研究给出一种基于研究主题的文献影响力评价方法。[方法/过程]以 Web of Science 数据库中 2011 年 - 2015 年间情报学领域 500 篇高被引文献为样本,利用 LDA 模型对样本文献进行主题建模,将主题对文献的支持度与文献被引频次相结合,计算特定主题文献的被引频次(specific topic cited frequency,简称 *STCF*),并根据每篇文献在相应主题内的 *STCF* 值对文献进行影响力排序。[结果/结论]结果表明,*STCF* 值能反映文献的主题内容、细粒度体现文献的学术地位、呈现文献研究主题的多元性,能够有效弥补被引频次、Altmetrics 指标的不足。

关键词: 被引频次 文献影响力 LDA 模型 *STCF* 值
分类号: G250.2
DOI: 10.13266/j.issn.0252 - 3116.2018.20.010

1 引言

定量评估学术文献的影响力是科研活动的一个重要环节,有助于评价个人、机构和国家的科研产出水平,可发现有价值的学术文献,满足相关人员的文献需求。对文献影响力的研究,国内外学者从多个角度进行了探讨,最常用的方法是考察文献的被引频次。国外学者如 C. C. Kam 等[1]使用阈值引文分析法,选取营销学研究中被引频次高于 18 的文献作为重点研究对象,确定了影响力最高的文献。然而,引用只是文献利用行为中的很小一部分[2],单纯依靠被引频次评估学术文献的影响力是不全面的。之后,相继出现了一些对文献影响力评估的新方法,如影响因子[3]、h 指数[4]、g 指数[5]、e 指数[6]、hg 指数[7]、文献时序排名算法(PTRA)[8]等,上述指标均是对被引频次指标的改进和完善。随着网络化、数字化技术的发展,不少学者如 P. Chen[9]、N. Ma[10]分别在被引频次的基础上利用 PageRank 算法实现了基于文献引用网络的文献排名;M. Krapivin 等[11]探讨了被引频次、h 指数和 PageRank 3 种评价方法在科学引文网络中的意义和影响,结果表明,利用不同方法对文献进行排名,会产生显著的差异。随着学术成果越来越多地在开放存取数字网络平台上发表,研究人员积极在各种学术社区中进行形式多样的学术交流,如评论、推荐、标注、转发、下载等,于是产生了形式多样的基于社会化网络的 Altmetrics 文献评价指标。T. Kortelainen 等[12]通过对社交媒体工具的分析,认为推荐、评论、链接分享等行为能够提高科研文献对普通大众的可见度,可将这些指标纳入到文献的效用评估中。Altmetrics 评价指标的兴起说明人们乐意从同行所在的虚拟学习社区中获取有价值的文献,不同的虚拟学习社区代表不同的学科方向和研究主题,可以说 Altmetrics 评价指标为在研究主题层面实现对文献影响力的评估提供了可借鉴的思路。

国内学者在引文频次的基础上也开展了相应的改进工作,如金碧辉提出的 A 指数[13]、AR 指数[14],吴强[15]提出的 W 指数,肖学斌[16]提出的 x 指数,韩毅[17]提出的 Pt 指数等。另有贾宁[18]通过对文献被引频次的年代分布进行分析,给出评价专著、期刊文章和学位论文 3 种文献类型的合理时间,创造性地提出“高位段持续时间和峰值高低”可以作为评价文献价值的参考指标;汪志伟等[19]在综合考虑多种因素的前提下,提出一种多维文献检索排序法,该方法以加权的方式

* 本文系国家社会科学基金项目“文献内容分析与引文分析融合的知识挖掘与发现研究”(项目编号:16BTQ074)研究成果之一。
作者简介:李秀霞(ORCID:0000 - 0002 - 3492 - 4768),教授,硕士生导师,E-mail:zyshao@126.com;宋凯(ORCID:0000 - 0001 - 7014 - 8498),硕士研究生。
收稿日期:2018 - 03 - 04 修回日期:2018 - 05 - 24 本文起止页码:88 - 94 本文责任编辑:王传清

将文献价值量化，有效地改善了文献检索排序法的效果；李长玲等[20]将引用文献的质量融于引文分析中，提出基于 PageRank 的引文分析法。国内也有关于 Altmetrics 文献评价指标的相关研究，如翟晓芳[21]融合基于引用的传统指标与基于社会化网络的 Altmetrics 指标，提出了适应国内学术环境的综合计量模型。

分析发现，已有研究主要以被引频次作为评价文献影响力的指标，在改进的一些方法中，多是将时间因素、社会网络方法与被引频次结合来评价文献的影响力，鲜有考虑文献主题因素的相关研究。Altmetrics 指标是基于社交网络对学术研究进行分析和传播的新型计量方法，拥有庞大的在线用户基础，能够客观地反映文献的社会影响力，但相关研究仍未将文献研究主题的差异纳入学术文献影响力评价中。为弥补上述不足，本文给出一种基于特定研究主题的学术文献影响力评价方法。

2 将研究主题纳入学术文献影响力评价中的科学性与合理性

将研究主题纳入学术文献影响力评价中，能够在主题分类基础上细粒度地测度学术文献的影响力，使学术文献影响力的评价结果更加科学、合理。

就评价客体（文献本身）来讲，不同学科领域的文献其研究对象、研究内容、研究方法、研究工具、表现形式等均存在较大的差异，而且研究人员数量不同，受众范围也不同，所以，不同学科领域的文献不能按相同的标准来衡量其影响力。“中文社会科学引文索引（CSSCI）”就是在学科分类基础上对期刊进行的排序，学术期刊是学术文献的载体，因此，对学术文献影响力的评价不能脱离学科分类独立进行。宋丽萍等[22]在探讨同行评议、影响计量学以及传统文献计量指标在科学评价中的有效性时也得到“科学评价中自然科学、社会科学具有较大差异”的结论。类似地，属于同一学科领域的文献（如同属情报学学科的文献，一篇研究数据挖掘技术方法的文献与一篇研究情报评价的文献），也不能按统一标准来评价其影响力。只有同一学科领域内相同研究主题的文献才有可比性，才能用统一的标准来评价。

就评价主体（评价者）来讲，研究方向相同的人一般评价熟悉的或相关的研究主题。同行评议、同行评审一直以来就是国内外对期刊、文献、机构等进行评价的一个流程，其中“同行”是指具有共同的追求目标、并由专家组成的“科学共同体”[23]；J. Liu 等[24]指出：要迅速获得科学文献对某特定领域用户的影响力，需通过分析该领域较活跃用户在网络社区中的各种讨论和交流活动来评估相关文献在该领域的影响力，因为具有相同研究方向的人才会在一起讨论、交流共同关注的话题，进而做出一定的评判。

可见，学术文献的影响力与学科领域、研究主题密切相关，将学科主题纳入文献影响力评价研究是科学的、合理的，也是必须的。

3 *STCF* 评价方法简介

本方法首先提取文献的主题，得到每个主题对一篇文献的支持度，之后将支持度与文献的被引频次相结合，得到一篇文献在特定主题上的被引频次，记为这篇文献的 *STCF* 值。

计算主要分 3 个步骤：

（1）主题建模。选取 LDA 模型抽取文献集中包含的主题，产生包含 T 个主题的文献－主题概率矩阵，矩阵中每一行表示 T 个主题在一篇文献中的概率分布，每一列表示某一个主题在 n 篇文献中的概率分布。一个主题在一篇文献中出现的概率大小称为该主题对这篇文献的支持度。一篇文献可以对应多个研究主题，各研究主题对同一篇文献的支持度各不相同，支持度越大，这篇文献与该主题的相关度越高。

（2）计算特定主题上文献的被引频次。① 对任意一篇文献 $P_i(i=1,2,\cdots n)$，查询 P_i 当前的被引频次，记为 C_i。② 对任意一个主题 $T_j(j=1,2,\cdots,T)$，确定 T_j 对 P_i 的支持度，记为 TS_{ij}；其值为小于 1 的百分数。③ 计算在主题 T_j 上文献 P_i 的被引频次 *STCF*。计算公式为：$STCF_i = N \times C_i \times TS_{ij}$。因为支持度 TS_{ij} 对被引频次 C_i 做了缩小的变化，降低了 *STCF* 值的敏感度，为提高其敏感性，前面加了一个敏感系数 N，N 的取值范围在 1－10 之间，具体取值根据所有 TS_{ij} 平均值的倒数大小来确定。

（3）特定主题文献影响力排名。在每个主题内，根据 *STCF* 值的大小对文献进行排名，得到与某个主题最相关的文献。一篇文献的 *STCF* 值越高，它在该主题内的影响力就越大，由此评价文献在不同主题内的影响力。

4 实证研究与结果分析

4.1 数据来源与处理

本研究实验数据来源于 Web of Science 核心合集数据库，Web of Science 数据库中共有 80 余种图书馆

与信息科学期刊，根据期刊内容，从中选取 2011 年 - 2015 年间情报学领域发文量最高的 6 种期刊（*Scientometrics*、*Journal of the American Medical Informatics Association*、*Journal of the American Society for Information Science and Technology*、*Information Procescing Management*、*International Journal of Information Management*、*Journal of Informetrics*），以其上被引频次最高的 500 篇文献为研究对象。下载每篇文献的标题与摘要信息作为实验数据集，并按被引频次对文献进行编号，分别标记 1 - 500，以方便计算过程中对文献的识别。接着对实验数据集进行预处理，使用 EnStemmer 工具实现去除停用词、词干化等自然语言处理规范化过程；之后将数据导入到 Excel 中，对每篇文献进行单词去重，同时删除高频出现但对本文没有研究意义的词语，如“advice”“journal”“record”“task”等，最终获得实验用的文本语料库。

4.2 *STCF* 值计算

在利用 LDA 模型对实验语料库进行主题建模时，依据文献[25]提出的主题之间的平均相似度来确定主题数目。实验发现，当主题数 T 设为 7 时，主题结构的平均相似度最小，此时对应的模型最优，因此，确定 500 篇样本文献涵盖 7 个研究主题。通过文献主题提取，形成文献 - 主题矩阵，见表 1。由文献 - 主题矩阵确定每个主题 T_j 对一篇文献 P_i 的支持度 TS_{ij}。

表 1 文献 - 主题矩阵（部分数据）

文献编号	文献篇名	主题 1 的 TS_{ij}	主题 2 的 TS_{ij}	主题 3 的 TS_{ij}	主题 4 的 TS_{ij}	主题 5 的 TS_{ij}	主题 6 的 TS_{ij}	主题 7 的 TS_{ij}
1	Negative results are disappearing from most disciplines and countries	0.097744361	0.082706767	0.165413534	0.090225564	0.112781955	0.180451128	0.067669173
2	Sentiment in Twitter Events	0.058823529	0.044117647	0.110294118	0.044117647	0.279411765	0.036764706	0.088235294
3	Sentiment Strength Detection for the Social Web	0.043209877	0.179012346	0.043209877	0.098765432	0.314814815	0.049382716	0.092592593
4	Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake	0.04	0.048	0.056	0.096	0.048	0.152	0.344
5	Towards a new crown indicator: Some theoretical considerations	0.327102804	0.046728972	0.056074766	0.14953271	0.046728972	0.130841121	0.056074766
6	2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text	0.057553957	0.503597122	0.035971223	0.057553957	0.035971223	0.043165468	0.079136691
7	Taiwan's National Health Insurance Research Database: administrative health care database as study object in bibliometrics	0.044117647	0.066176471	0.102941176	0.110294118	0.125	0.279411765	0.058823529
8	The Leiden ranking 2011/2012: Data collection, indicators, and interpretation	0.404411765	0.044117647	0.044117647	0.088235294	0.044117647	0.125	0.051470588
9	Impact factor: outdated artefact or stepping-stone to journal certification?	0.148148148	0.092592593	0.111111111	0.101851852	0.083333333	0.12037037	0.064814815
10	Social disparities in internet patient portal use in diabetes: evidence that the digital divide extends beyond access	0.041666667	0.041666667	0.075	0.058333333	0.108333333	0.075	0.1
11	An empirical investigation of mobile banking adoption: The effect of innovation attributes and knowledge-based trust	0.033557047	0.040268456	0.483221477	0.046979866	0.046979866	0.060402685	0.087248322
12	Turning the Tables on Citation Analysis One More Time: Principles for Comparing Sets of Documents	0.409836066	0.081967213	0.040983607	0.057377049	0.06557377	0.090163934	0.049180328
13	Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study	0.04787234	0.138297872	0.031914894	0.319148936	0.031914894	0.117021277	0.042553191
14	Beyond the hype: Big data concepts, methods, and analytics	0.046783626	0.099415205	0.035087719	0.333333333	0.099415205	0.058479532	0.087719298
15	Social media competitive analysis and text mining: A case study in the pizza industry	0.075471698	0.106918239	0.037735849	0.06918239	0.062893082	0.037735849	0.433962264
16	Next-generation phenotyping of electronic health records	0.06741573	0.112359551	0.08988764	0.168539326	0.08988764	0.101123596	0.08988764
17	To be or not to be in social media: How brand loyalty is affected by social media?	0.046666667	0.04	0.52	0.046666667	0.033333333	0.04	0.16

根据各文献在不同主题上的支持度 TS_{ij} 平均值的倒数（在 2.143 - 4.903 之间）确定敏感系数，以最小数 2.143 的临近整数为选取敏感系数的原则，取 $N=2$。然后，根据与每个主题最相关的词对主题赋予标签，7 个研究主题分别为情报评价、医学信息分析、电子商务与决策支持、开放数据、文献计量、新媒体研究、社会网络分析等。

在 Web of Science 数据库查询文献 P_i 的被引频次 C_i，由 *STCF* 的计算公式计算每篇文献的 *STCF* 值，据此对文献进行影响力排名。为了说明 *STCF* 值在文献影响力评价中的优势，将 *STCF* 值与文献的总被引频次、Altmetrics 指标值进行对比。刘晓娟等[26]在对多个 Altmetrics 指标进行分析时，发现 Mendeley 和 Twitter 对图书情报领域论文的评价更有参考价值。因此，本研究以在 Mendeley 平台的注册用户将该篇文章加入到 Mylibary 的人数作为 Altmetrics 指标值。对比结果见表 2。

4.3 结果分析

4.3.1 *STCF* 值与被引频次、Altmetrics 值的相关性 在 7 个主题上将 *STCF* 值与被引频次、Altmetrics 做相关分析，共得到 7 组 *STCF* 值与被引频次、Altmetrics 的相关系数，见表 3。

由表 3 可以看出，在多数主题上，*STCF* 值与被引频次的相关系数在 0.496 - 0.619 之间，在个别主题上两者的相关系数较低（在主题 6 上为 -0.092），但总体来看，*STCF* 值与被引频次的相关度较高，说明 *STCF* 值评价结果与基于引用关系的评价结果具有较高的一致性，这是因为 *STCF* 值并未完全否定被引频次对文献影响力的作用，而是把被引频次看作是 *STCF* 值计算公式中的一个因子，是在肯定文献学术价值的基础上增加了社会关注度因素。*STCF* 值与 Altmetrics 评价值的相关性较低，在多数主题上，*STCF* 值与 Altmetrics 的相关系数在 0.125 - 0.370 之间，在个别主题（如主题 6）上，两者的相关系数为 -0.092。这说明利用不同方法对文献进行影响力评价，其评价结果会存在较大的差异；同时，*STCF* 值虽然同时考虑了文献的学术价值和社会关注度，但相对而言，更偏重反映文献的学术价值。

4.3.2 *STCF* 值的优势 与被引频次和 Altmetrics 指标相比，在学术文献影响力评价中，*STCF* 值具有以下明显优势：

表 2　高影响力文献的 *STCF* 值与被引频次、Altmetrics（将该篇文章加入到 Mylibary 的人数）的对比（部分数据）

主题	文献编号	题名	*STCF*	被引频次	Altmetrics
1 情报评价	1－1	Towards a new crown indicator: Some theoretical considerations	98.13	150	135
	1－2	The Leiden ranking 2011/2012: Data collection, indicators, and interpretation	93.82	116	125
	1－3	Turning the Tables on Citation Analysis One More Time: Principles for Comparing Sets of Documents	76.23	93	44
	1－4	The skewness of science in 219 sub-fields and a number of aggregates	54.71	63	40
	1－5	How Fractional Counting of Citations Affects the Impact Factor: Normalization in Terms of Differences in Citation Potentials Among Fields of Science	47.76	56	62
2 医学信息分析	2－1	2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text	133.96	133	223
	2－2	Sentiment Strength Detection for the Social Web	58.72	164	372
	2－3	A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries	54.04	52	128
	2－4	Portability of an algorithm to identify rheumatoid arthritis in electronic health records	39.29	68	21
	2－5	Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010	38.46	46	138
3 电子商务与决策支持	3－1	An empirical investigation of mobile banking adoption: The effect of innovation attributes and knowledge-based trust	94.71	98	391
	3－2	To be or not to be in social media: How brand loyalty is affected by social media?	84.24	81	925
	3－3	The effects of relationship quality and switching barriers on customer loyalty	78.37	72	240
	3－4	Negative results are disappearing from most disciplines and countries	57.89	175	453
	3－5	The impact of consumer trust on attitudinal loyalty and purchase intentions in B2C e-marketplaces: Intermediary trust vs. seller trust	56.00	56	346
4 开放数据	4－1	Beyond the hype: Big data concepts, methods, and analytics	60	90	2 200
	4－2	Validation of a common data model for active safety surveillance research	59.93	58	38
	4－3	Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study	58.72	92	33
	4－4	Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the e-MERGE Network experience	48.86	48	140
	4－5	Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the e-MERGE network	45.14	79	183
5 文献计量	5－1	Sentiment Strength Detection for the Social Web	103.26	164	372
	5－2	Sentiment in Twitter Events	96.68	173	357
	5－3	A preliminary test of Google Scholar as a source for citation data: a longitudinal study of Nobel prize winners	47.47	55	73
	5－4	A study of open access journals using article processing charges	41.14	48	133
	5－5	Negative results are disappearing from most disciplines and countries	39.47	175	453
6 新媒体研究	6－1	Taiwan's National Health Insurance Research Database: administrative health care database as study object in bibliometrics	68.74	123	40
	6－2	Negative results are disappearing from most disciplines and countries	63.16	175	453
	6－3	Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake	46.82	154	649
	6－4	A Heuristic Approach to Author Name Disambiguation in Bibliometrics Databases for Large-Scale Research Assessments	43.12	78	48
	6－5	Towards a new crown indicator: Some theoretical considerations	39.25	150	135
7 社会网络分析	7－1	Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake	105.95	154	649
	7－2	Social media competitive analysis and text mining: A case study in the pizza industry	78.11	90	725
	7－3	Cloud computing as an innovation: Percepetion, attitude, and adoption	54.25	62	457
	7－4	Reaching for the cloud: How SMEs can manage	44.67	67	432
	7－5	The usage and adoption of cloud computing by small and medium businesses	39.27	66	634

注：表中各个主题下的文献按 STCF 值由大到小排序

（1）*STCF* 值能够从主题内容上反映文献的学术价值。文献被引频次单纯从被引用的角度来评价文献的影响力，虽然反映了文献的学术影响力和学术价值，但无法体现文献的研究主题和内容；Altmetrics 指标主要是从文献的读者数以及读者通过在线社交媒体进行交流过程中的推荐、评论等交互行为和传统媒介的社会传播深度来度量，偏向社会影响力和社会关注度，虽然社群的特征能够反映其关注的主题内容，但较少涉

表 3 7 个主题中 500 篇文献的 *STCF* 值与被引频次的相关系数

	STCF 值与被引频次	*STCF* 值与 Altmetrics
主题 1	.496 **	.124 **
主题 2	.606 **	.207 **
主题 3	.564	.325
主题 4	.609 **	.232 **
主题 5	.516	.256
主题 6	-.092 *	-.150 **
主题 7	.616	.370

说明：**表示在置信度（双测）为 0.01 时，相关性是显著的；* 表示在置信度（双测）为 0.05 时，相关性是显著的

及与论文质量相关的学术价值[27]。而 *STCF* 值是研究主题对文献的支持度 TS_{ij} 与文献的被引频次 C_i 的融合，研究主题对文献的支持度反映了文献的主题内容，被引频次体现了文献的学术价值，因此，*STCF* 值从主题内容上反映了文献的学术价值。

（2）*STCF* 值能够在同类研究主题内反映文献的学术地位。学术文献的被引频次和 Altmetrics 指标值最直接的体现应该是在文献对应研究主题上研究力量的强弱，因为不同研究方向的研究人员和爱好者数量不同，即使学术水平相同的文献，其被引频次、Altmetrics 指标值也会存在差异；反之，被引频次、Altmetrics 指标值都相同的文献，其学术水平也不一定等同。而利用 *STCF* 值可以发现那些研究力量相对薄弱、社会影响力不是很高、却具有较高学术水平的文献。比如表 2 中第 4 类“开放数据”中文献 4 - 1 的 *STCF* 值为 60，小于第 1 类“情报评价”研究主题下文献 1 - 3 对应的 *STCF* 值（76.23）；第 6 类“新媒体研究”中文献 6 - 1 的 *STCF* 值为 68.74，小于第 3 类“电子商务与决策支持”研究主题下文献 3 - 3 对应的 *STCF* 值（78.37）。上述 *STCF* 值低的文献在主题 4、主题 6 中的排名都是第一位，而 *STCF* 值高的文献在主题 1、主题 3 中的排名却是第 3 位。可见，*STCF* 值是在同类研究主题内比较文献的影响力，能够在主题内容上反映学术文献的合理地位。

（3）*STCF* 值体现了文献研究主题的多元性和倾向性。一篇文献往往会有不同的研究主题，在进行主题分类时同一篇文献会被分到不同的主题类团中。根据表 2 的数据以及被引频次、Altmetrics 指标的定义，即使先将文献按主题分类，一篇文献无论被分到哪个主题类团中，其被引频次、Altmetrics 指标值都分别对应同一个值，如文献 3 - 4“Negative results are disappearing from most disciplines and countries”的被引频次为 175，Altmetrics 指标值为 453。在所有主题中，若按被引频次排名，该文献排名都是第一位；若按 Altmetrics 指标排名，该文献都是排在第 11 位。但一篇文献在研究主题上总有轻、重之别，在其涉及的所有研究主题上的评价值都相同是不符合实际的。*STCF* 值是根据文献在不同研究主题上的支持度得到的，同一篇文献，对不同主题的支持度有别，因此，在不同主题内其学术地位就不同，如上述文献在主题 1 - 7 中的排名分别是第 15、第 7、第 4、第 11、第 5、第 2、第 15，详见如表 4 所示：

表 4 文献 3 - 4 在不同主题内的评价结果对比

文献 3 - 4	被引频次（排序）	Altmetrics（排序）	*SCTF* 值（排序）
主题 1	175（1）	453（11）	34.22（15）
主题 2	175（1）	453（11）	28.94（7）
主题 3	175（1）	453（11）	57.89（4）
主题 4	175（1）	453（11）	31.58（11）
主题 5	175（1）	453（11）	39.47（5）
主题 6	175（1）	453（11）	63.16（2）
主题 7	175（1）	453（11）	23.68（15）

可见，*STCF* 值不仅体现了学术文献研究主题的多元性，而且还能反映一篇文献在不同研究主题上着力大小的差异。

5 结论

本研究提出一种基于研究主题的学术文献影响力评价新指标，即 *STCF* 值，并以情报学领域 500 篇高被引文献为研究样本进行了实证研究，通过对比发现 *STCF* 值具有其独特的优势。

本研究的主要结论如下：①利用 LDA 模型对 500 篇高被引文献进行主题提取，发现 500 篇文献涵盖 7 个研究主题：情报评价、医学信息分析、电子商务与决策支持、开放数据、文献计量、新媒体研究、社会网络分析等。②计算每篇文献在各主题内的 *STCF* 值，并在不同研究主题内对文献进行影响力排序。通过与被引频次、Altmetrics 评价值对比，发现：*STCF* 值与被引频次相关度较高，与 Altmetrics 的相关度较低，说明 *STCF* 值虽然同时考虑了文献的学术地位和社会关注度，但相对而言，更偏重反映文献的学术价值。③通过 3 种评价方法的优势对比，发现：基于主题分类的 *STCF* 值评价方法能够同时从主题内容和社会关注度上反映文献的学术价值和学术地位，符合同行评议的评判规则；*STCF* 值同时反映了文献研究主题的多元化属性。可以说，本研究提出的 *STCF* 值是对传统文献被引频次

评价方法的改进和完善,为挖掘不同研究主题内具有高影响力和引用价值的文献提供了新的视角和途径。

STCF 值的不足之处在于:有些文献(如综述性的文献)涉及的主题较多,如果在每个研究主题上支持度都不高,那么在各个主题上的 *STCF* 值就会相对偏低,这种文献的影响力可能会被低估。另外,文献的被引次数会不断变化,而且不同学科领域文献的半衰期也不同,时间变量是评价学科主题内文献影响力的一个重要因素,本研究对时间变量的忽视也会导致评价结果的偏颇。

在未来研究中,将考虑对所有主题进行主成分提取,并将成分因子的方差贡献率占累计贡献率的比重对各主成分因子赋权,再累计各主成分加权后对文献的支持度,在此基础上与文献被引频次融合,同时,将时间变量引入主题影响力评价体系中,以更加客观、公正地评价学术文献的影响力。

参考文献:

[1] KAM C C, PIKKI L, KARTONO L. A threshold citation analysis in marketing research[J]. European journal of marketing,2012,46(1):134－156.

[2] 刘晓娟,周建华,尤斌. 基于 Mendeley 与 WoS 的选择性计量指标与传统科学计量指标相关性研究[J]. 图书情报工作,2015,59(3): 112－118.

[3] 许琦. 一种基于引证网络的文献影响因子计算方法[J],情报理论与实践,2011,34(7):98－102.

[4] HIRSCH J E. An index to quantify an individual,s research output [J]. Proceedings of National Academy of Sciences, 2005,102(46):16569 －16572.

[5] EGGHE L. An improvement of the h-index: the g-index[J]. ISSI newsletter,2006,2(1):8－9.

[6] ZHANG C. The e-index,complementing the h-index for excess citations[J]. Plos One,2009,4(5) : 1－4.

[7] ALONSO S,CABRERIZO F J,HERRERA-VIEDMA E,et al. Hg-index: a new index to characterize the scientific output of researchers based on the h-and g -indices[J]. Scientometrics,2010 , 82(2) :391－400.

[8] MUSHTAQ A H, LU S F, BASHEER A. Scientific research paper ranking algorithm PTRA: a tradeoff between time and citation network[J]. Applied mechanics and materials,2014, 551 (8) :603－611.

[9] CHEN P, XIE H, MASLOV S, et al. Finding scientific gems with Google's PageRank algorithm[J]. Journal of informetrics,2007,1(1):8－15.

[10] MA N. Bringing PageRank to the citation analysis[J]. Information processing & management, 2008,44(2): 800－810.

[11] KRAPIVIN M, MARCHESE M, CASATI F. Exploring and Understanding Scientific Metrics in Citation Networks[J]. Complex sciences, 2009(5):1550－1563.

[12] KORTELAINEN T. "Everything is plentiful-except attention". Attention data of scientific journals on social web tools[J]. Journal of informetrics, 2012 , 6(4): 661－668.

[13] JIN B H, LIANG L M, ROUSSEAU R, et al. The R-and AR-indices: complementing the h-index[J]. Chinese science bulletin, 2007, 52(6): 855－863.

[14] 金碧辉, ROUSSEAU R. R 指数、AR 指数:h 指数功能扩展的补充指标[J]. 科学观察, 2007, 2(3):1－8.

[15] WU Q. The w index: a measure to assess scientific impact by focusing on widely cited papers[J]. Journal of the American Society for Information Science and Technology,2010,61(3):690－614.

[16] 肖学斌. x 指数:描述研究人员论文水献计量新指数[J]. 图书情报知识,2015 (2): 93－99.

[17] 韩毅,夏慧. 时间因素视角下科研人员评价的 Pt 指数研究[J]. 中国图书馆学报, 2015,41(6): 73－85.

[18] 贾宁. 文献被引的年代分布对被引文献评价的意义——以物理学科为例[J]. 图书馆杂志,2016,35 (12):55－62.

[19] 汪志伟,邹艳妮,吴舒霞. PageRank 算法应用在文献检索排序中的研究及改进[J]. 情报理论与实践,2016(11):126－130,144.

[20] 李长玲,翟雪梅. 基于 PageRank 的引文分析方法探讨[J]. 情报理论与实践,2007,30(1):122－124.

[21] 翟晓芳,刘全明,程耀东,等. 结合社会化网络的文献综合计量模型[J]. 计算机工程,2016, 42(6):21－26.

[22] 宋丽萍 ,王建芳 ,王树义. 科学评价视角下 F1000、Mendeley 与传统文献计量指标的比较[J]. 中国图书馆学报 , 2014, 40(4):48－54.

[23] 王国豫,朱晓林. 同行评议与"外行"评议[J]. 科学学研究, 2015, 33(8):1121－1126,1133.

[24] LIU J,ADIE E. New perspectives on article-level metrics: developing ways to assess research uptake and impact online[J]. Insights,2013,27(2) : 153－158.

[25] 曹娟,张勇东,李锦涛,等. 一种基于密度的自适应最优 LDA 模型选择方法[J]. 计算机学报,2008, 31(10):1780－1787.

[26] 刘晓娟,宰冰欣. 图书情报领域文献的 Altmetrics 指标分析[J]. 图书情报工作,2015,59(18):108－116.

[27] 赵蓉英,郭凤娇,谭洁. 基于 Altmetrics 的学术论文影响力评价研究——以汉语言文学学科为例[J]. 中国图书馆学报, 2016, 42(1):96－108.

作者贡献说明:

李秀霞:研究思路、研究方法设计,论文撰写及修改;

宋凯:数据收集及程序调试。

STCF Value: A New Index to Evaluate the Academic Literature Influence Based on Research Topic

Li Xiuxia Song Kai

School of Communication, Qufu Normal University, Rizhao 276826

Abstract: [**Purpose/significance**] The indexes of evaluating literature influence continue to innovate, but a lot of indexes are lack of assessment of the literature influence on the topic level. Aimed to solve this problem, this paper presents a method based on the research topic. [**Method/process**] This paper chooses 500 highly cited literature as samples which attribute to the field of information science in the Web of Science database. LDA model is used to model the literature subject. The support of the topic of the literature is combined with the cited frequency to compute the specific topic cited frequency(*STCF*). We rank the literature according to the *STCF* value of each document in the corresponding subject. [**Result/conclusion**] The empirical research shows that *STCF* value reflects the subject content, the subject diversity, and the academic status of literature. This method effectively compensates for the shortage of the cited frequency and Altmetrics index to evaluate the literature impact.

Keywords: citation frequency literature influence LDA model *STCF* value

关于在学术论文署名中常见问题或错误的诚信提醒

恪守科研道德是从事科技工作的基本准则,是履行党和人民所赋予的科技创新使命的基本要求。中国科学院科研道德委员会办公室根据日常科研不端行为举报中发现的突出问题,总结当前学术论文署名中的常见问题和错误,予以提醒,倡导在科研实践中的诚实守信行为,努力营造良好的科研生态。

提醒一:论文署名不完整或者夹带署名。应遵循学术惯例和期刊要求,坚持对参与科研实践过程并做出实质性贡献的学者进行署名,反对进行荣誉性、馈赠性和利益交换性署名。 提醒二:论文署名排序不当。按照学术发表惯例或期刊要求,体现作者对论文贡献程度,由论文作者共同确定署名顺序。反对在同行评议后、论文发表前,任意修改署名顺序。部分学科领域不采取以贡献度确定署名排序的,从其规定。

提醒三:第一作者或通讯作者数量过多。应依据作者的实质性贡献进行署名,避免第一作者或通讯作者数量过多,在同行中产生歧义。

提醒四:冒用作者署名。在学者不知情的情况下,冒用其姓名作为署名作者。论文发表前应让每一位作者知情同意,每一位作者应对论文发表具有知情权,并认可论文的基本学术观点。

提醒五:未利用标注等手段,声明应该公开的相关利益冲突问题。应根据国际惯例和相关标准,提供利益冲突的公开声明。如资金资助来源和研究内容是否存在利益关联等。

提醒六:未充分使用志(致)谢方式表现其他参与科研工作人员的贡献,造成知识产权纠纷和科研道德纠纷。

提醒七:未正确署名所属机构。作者机构的署名应为论文工作主要完成机构的名称,反对因作者所属机构变化,而不恰当地使用变更后的机构名称。

提醒八:作者不使用其所属单位的联系方式作为自己的联系方式。不建议使用公众邮箱等社会通讯方式作为作者的联系方式。

提醒九:未引用重要文献。作者应全面系统了解本科研工作的前人工作基础和直接相关的重要文献,并确信对本领域代表性文献没有遗漏。

提醒十:在论文发表后,如果发现文章的缺陷或相关研究过程中有违背科研规范的行为,作者应主动声明更正或要求撤回稿件。

院属各单位应根据以上提醒,结合本单位学科特点和学术惯例,对科研人员进行必要的教育培训,让每一位科研工作者对学术论文署名保持高度的责任心,珍惜学术荣誉、抵制学术不端行为,将科研诚信贯穿于学术生涯始终。

来源:中国科学院监督与审计局